Sparse coding for multitask and transfer learning

Andreas Maurer[(1)], **Massimiliano Pontil**[(2)], **Bernardino Romera-Paredes**[(3)]

(1) Adalbertstr. 55, D-80799, München, GERMANY
E-mail: *am@andreas-maurer.eu*

(2) Department of Computer Science
Centre for Computational Statistics and Machine Learning
University College London
Gower Street, London WC1E, England, UK
E-mail: *m.pontil@cs.ucl.ac.uk*

(3) Department of Computer Science
UCL Interactive Centre
University College London
Gower Street, London WC1E, England, UK
E-mail: *bernardino.paredes.09@ucl.ac.uk*

Abstract

We present an extension of sparse coding to the problems of multitask and transfer learning. The central assumption of the method is that the tasks parameters are well approximated by sparse linear combinations of the atoms of a dictionary on a high or infinite dimensional space. This assumption, together with the large quantity of available data in the multitask and transfer learning settings, allows a principled choice of the dictionary. We provide bounds on the generalization error of this approach, for both settings. Preliminary experiments indicate the advantage of the sparse multitask coding method over single task learning and a previous method based on orthogonal and dense representation of the tasks.

1 Introduction

The last decade has witnessed many efforts of the machine learning community to exploit assumptions of sparsity in the design of algorithms. A central development in this respect is the Lasso [29], which estimates a linear predictor in a high dimensional space under a regularizing ℓ_1-penalty. Theoretical results guarantee a good performance of this method under the assumption that the vector corresponding to the underlying predictor is sparse, or at least has a very small ℓ_1-norm, see for example [11, 12, 31] and references therein.

In this work, we consider the case where the predictors are linear combinations of the atoms of a dictionary of linear functions on a high or infinite dimensional space, and we assume that we are free to choose the dictionary. We will show that a principled choice is possible, if there are many learning problems, or "tasks", and there exists a dictionary allowing sparse, or nearly sparse representations of all or most of the underlying predictors. In such a case we can then exploit the larger quantity of available data to estimate the "good" dictionary and still reap the benefits of the Lasso for the individual tasks. This paper gives theoretical and experimental justification of this claim, both in the domain of multitask learning, where the new representation is applied to the tasks from which it was generated, and in the domain of learning to learn, where the dictionary is applied to new tasks of the same environment.

Our work combines ideas from sparse coding [25, 26], multitask learning [1, 3, 10, 13, 14] and learning to learn [7, 30]. There is a vast literature on these subjects and the list of papers provided here is necessarily incomplete. Learning to learn (also called inductive bias learning or transfer learning) has been proposed by Baxter [7] and an error analysis is provided therein, showing that a common representation which performs well on the training tasks will also generalize to new tasks obtained from the same "environment". The precursors of the analysis presented here are [23] and [22]. The first paper provides a bound on the reconstruction error of sparse coding and may be seen as a special case of the ideas presented here in the case of infinite sample size. The second paper provides a learning to learn analysis of the multitask feature learning method in [3]. There are other works such as [19, 27] which have explored the application of sparse coding for supervised learning. The main idea pursed in those papers is to simultaneously learn a dictionary from the input data and at the same time use the coding vectors as features for a supervised learning algorithm. Such features are a non-linear transformation of the input and the tasks are assumed to be related because they share the same dictionary used to produce this representation. In our approach, we seek a dictionary which represents well the tasks' regression vectors, assuming that these are a sparse combination of the dictionary elements. In other words, the feature learned by our method are a linear transformation of the input data and sparsity is enforced on the tasks' regression coefficients. We note that at the time of the paper writing a method very similar to ours has been proposed for multitask learning [17]. Here we present a probabilistic analysis which complements well with the practical insights in [17], highlight the connection to sparse coding [25] and address the different problem of learning to learn.

The paper is organized in the following manner. In Section 2, we set up our notation and introduce the learning problem. In Section 3, we present our learning bounds for multitask learning and learning to learn. In Section 4 we report on numerical experiments. Section 5 contains concluding remarks.

2 Method

In this section, we turn to a technical exposition of the proposed method, introducing some necessary notation on the way.

Let H be a finite or infinite dimensional Hilbert space with inner product $\langle \cdot, \cdot \rangle$, norm $\|\cdot\|$, and fix an integer K. We study the problem

$$\min_{D \in \mathcal{D}_K} \frac{1}{T} \sum_{t=1}^{T} \min_{\gamma \in \mathcal{C}_\alpha} \frac{1}{m} \sum_{i=1}^{m} \ell \left(\langle D\gamma, x_{ti} \rangle, y_{ti} \right), \tag{2.1}$$

where

- $\mathcal{D}_K$ is the set of K-dictionaries (or simply dictionaries), which means that every $D \in \mathcal{D}_K$ is a linear map $D : \mathbb{R}^K \to H$, such that $\|De_k\| \leq 1$ for every one of the canonical basis vectors e_k of $\mathbb{R}^K$. The number K can be regarded as one of the regularization parameters of our method.

- $\mathcal{C}_\alpha$ is the set of vectors γ in $\mathbb{R}^K$ satisfying $\|\gamma\|_1 \leq \alpha$. The ℓ_1-norm constraint implements the assumption of sparsity and α is the other regularization parameter. Different sets $\mathcal{C}_\alpha$ could be readily used in our method, such as those associated with ℓ_p-norms or mixed-norm, see e.g. [15].

- $\mathbf{Z} = ((x_{ti}, y_{ti}) : 1 \leq i \leq m, 1 \leq t \leq T)$ is a dataset on which our algorithm operates. Each $x_{ti} \in H$ represents an input vector, and y_{ti} is a corresponding real valued label. We also write $\mathbf{Z} = (\mathbf{X}, \mathbf{Y}) = (\mathbf{z}_1, \ldots, \mathbf{z}_T) = ((\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_T, \mathbf{y}_T))$ with $\mathbf{x}_t = (x_{t1}, \ldots, x_{tm})$ and $\mathbf{y}_t = (y_{t1}, \ldots, y_{tm})$. The index t identifies a learning task, and $\mathbf{z}_t$ are the corresponding training points, so the algorithm operates on T tasks, each of which is represented by m example pairs.

- ℓ is a loss function where $\ell(y, y')$ measures the loss incurred by predicting y when the true label is y'. We assume that ℓ has values in $[0, 1]$ and has Lipschitz constant L in the first argument for all values of the second argument.

The minimum in (2.1) is zero if the data is generated according to a noise-less model which postulates that there is a "true" dictionary $D^* \in \mathcal{D}_{K^*}$ with K^* atoms and vectors $\gamma_1^*, \ldots, \gamma_T^*$ satisfying $\|\gamma_t^*\|_1 \leq \alpha^*$, such that an input $x \in H$ generates the label $y = \langle D^* \gamma_t^*, x \rangle$ in the context of task t. If $K \geq K^*$ and $\alpha \geq \alpha^*$ then the minimum in (2.1) is zero. In Section 4, we will present experiments with such a generative model, when noise is added to the labels, that is $y = \langle D^* \gamma_t^*, x \rangle + \zeta$ with $\zeta \sim \mathcal{N}(0, \sigma)$, the standard normal distribution.

The method (2.1) should output a minimizing dictionary $D(\mathbf{Z}) \in \mathcal{D}_K$ as well as minimizing codes $\gamma_1(\mathbf{Z}), \ldots, \gamma_T(\mathbf{Z})$ corresponding to the different tasks. Our implementation, described below, does not guarantee exact minimization, because of the non-convexity of the problem. Below predictors are always linear, specified by a vector $w \in H$, predicting the label $\langle w, x \rangle$ for an input $x \in H$, and a learning algorithm is a rule which assigns a predictor $A(\mathbf{z})$ to a given data set $\mathbf{z} = ((x_i, y_i) : 1 \leq i \leq m) \in (H \times \mathbb{R})^m$.

We note that a method similar to (2.1) has been proposed in [17], where the Frobenius norm on the dictionary is used in place of the ℓ_2/ℓ_∞-norm employed here.

3 Learning bounds

In this section, we present learning bounds for method (2.1), both in the multitask learning and learning to learn settings, and discuss the special case of sparse coding.

3.1 Multitask learning

Let $\mu_1, \ldots, \mu_T$ be probability measures on $H \times \mathbb{R}$. We interpret $\mu_t(x, y)$ as the probability of observing the input/output pair (x, y) in the context of task t. For each of these tasks an i.i.d. training sample $\mathbf{z}_t = ((x_{ti}, y_{ti}) : 1 \leq i \leq m)$ is drawn from $(\mu_t)^m$ and the ensemble $\mathbf{Z} \sim \prod_{t=1}^T \mu_t^m$ is input to algorithm (2.1). Upon returning of a minimizing $D(\mathbf{Z})$ and $\gamma_1(\mathbf{Z}), \ldots, \gamma_T(\mathbf{Z})$, we will use the predictor $D(\mathbf{Z})\gamma_t(\mathbf{Z})$ on the t-th task. The average over all tasks of the expected error incurred by these predictors is

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}_{(x,y)\sim\mu_t}\left[\ell\left(\left\langle D(\mathbf{Z})\gamma_t(\mathbf{Z}), x\right\rangle, y\right)\right].$$

We compare this *task-average risk* to the minimal analogous risk obtainable by any dictionary D and any set of vectors $\gamma_1, \ldots, \gamma_T \in \mathcal{C}_\alpha$. Our first result is a bound on the excess risk.

Theorem 1. *Let $\delta > 0$ and let $\mu_1, \ldots, \mu_T$ be probability measures on $H \times \mathbb{R}$. With probability at least $1 - \delta$ in the draw of $\mathbf{Z} \sim \prod_{t=1}^T \mu_t^m$ we have*

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}_{(x,y)\sim\mu_t}\left[\ell\left(\left\langle D(\mathbf{Z})\gamma_t(\mathbf{Z}), x\right\rangle, y\right)\right] - \inf_{D\in\mathcal{D}_K}\frac{1}{T}\sum_{t=1}^T \inf_{\gamma\in\mathcal{C}_\alpha}\mathbb{E}_{(x,y)\sim\mu_t}\left[\ell\left(\left\langle D\gamma, x\right\rangle, y\right)\right]$$
$$\leq L\alpha\sqrt{\frac{2S_1(\mathbf{X})(K+12)}{mT}} + L\alpha\sqrt{\frac{8S_\infty(\mathbf{X})\ln(2K)}{m}} + \sqrt{\frac{8\ln 4/\delta}{mT}},$$

where $S_1(\mathbf{X}) = \frac{1}{T}\sum_{t=1}^T \operatorname{tr}\left(\hat{\Sigma}(\mathbf{x}_t)\right)$ and $S_\infty(\mathbf{X}) = \frac{1}{T}\sum_{t=1}^T \lambda_{\max}\left(\hat{\Sigma}(\mathbf{x}_t)\right)$. Here $\hat{\Sigma}(\mathbf{x}_t)$ is the empirical covariance of the input data for the t-th task, $\operatorname{tr}(\cdot)$ denotes the trace and $\lambda_{\max}(\cdot)$ the largest eigenvalue.

We state several implications of this theorem.

1. The quantity $S_1(\mathbf{X})$ appearing in the bound is just the average square norm of the input data points, while $S_\infty(\mathbf{X})$ is roughly the average inverse of the observed dimension of the data for each task. Suppose that $H = \mathbb{R}^d$ and that the data-distribution is uniform on the surface of the unit ball. Then $S_1(\mathbf{X}) = 1$ and for $m \ll d$ it follows from Levy's isoperimetric inequality (see e.g. [18]) that $S_\infty(\mathbf{X}) \approx 1/m$, so the corresponding term behaves like $\sqrt{\ln K}/m$. If the minimum in (2.1) is small and T is large enough for this term to become dominant then there is a significant advantage of the method over learning the tasks independently. If the data is essentially low dimensional, then $S_\infty(\mathbf{X})$ will be large, and in the extreme case, if the data is one-dimensional for all tasks then $S_\infty(\mathbf{X}) = S_1(\mathbf{X})$ and our bound will always be worse by a factor of $\ln K$ than standard bounds for independent single task learning as in [6]. This makes sense, because for low dimensional data there can be little advantage to multi-task learning.

2. In the regime $T < K$ the bound is dominated by the term of order $\sqrt{S_1(\mathbf{X})\,K/mT} > \sqrt{S_1(\mathbf{X})/m}$. This is easy to understand, because the dictionary atoms De_k can be chosen independently, separately for each task, so we could at best recover the usual bound for linear models and there is no benefit from multi-task learning.

3. Consider the noiseless generative model mentioned in Section 2. If $K \geq K^*$ and $\alpha \geq \alpha^*$ then the minimum in (2.1) is zero. In the bound the overestimation of K^* can be compensated by a proportional increase in the number of tasks considered and an only very minor increase of the sample size m, namely $m \to (\ln K^*/\ln K)\,m$.

4. Suppose that we concatenate two sets of tasks. If the tasks are generated by the generative model described in Section 2 then the resulting set of tasks is also generated by such a model, obtained by concatenating the lists of atoms of the two true dictionaries D_1^* and D_2^* to obtain the new dictionary D^* of length $K^* = K_1^* + K_2^*$ and taking the union of the set of generating vectors $\{\gamma_t^{*1}\}_{t=1}^T$ and $\{\gamma_t^{*2}\}_{t=1}^T$, extending them to $\mathbb{R}^{K_1^*+K_2^*}$ so that the supports of the first group are disjoint from the supports of the second group. If $T_1 = T_2$, $K_1^* = K_2^*$ and we train with the correct parameters, then the excess risk for the total task set increases only by the order of $1/\sqrt{m}$, independent of K, despite the fact that the tasks in the second group are in no way related to those in the first group. This is directly related to avoiding negative transfer. Negative transfer happens in situations when we attempt to "transfer knowledge" between unrelated tasks, thereby decreasing the statistical performance. The bound in Theorem 1 suggests that our method avoids negative transfer by implicitly finding the right clusters of mutually related tasks.

5. Consider the alternative method of subspace learning (SL) where $\mathcal{C}_\alpha$ is replaced by an euclidean ball of radius α. With similar methods one can prove a bound for SL where, apart from slightly different constants, $\sqrt{\ln K}$ above is replaced by K. SL will be successful and outperform the proposed method, whenever K can be chosen small, with $K < m$ and the vector γ_t^* utilize the entire span of the dictionary. For large values of K, a correspondingly large number of tasks and sparse γ_t^* the proposed method will be superior.

The proof of Theorem 1, which is given in Section B.1 of the supplementary appendix, uses standard methods of empirical process theory, but also employs a concentration result related to Talagrand's convex distance inequality to obtain the crucial dependence on $S_\infty(\mathbf{X})$. At the end of Section B.1 we sketch applications of the proof method to other regularization schemes, such as the one presented in [17].

3.2 Learning to learn

There is no absolute way to assess the quality of a learning algorithm. Algorithms may perform well on one kind of task, but poorly on another kind. It is important that an algorithm performs well on those tasks which it is likely to be applied to. To formalize this, Baxter [7] introduced the notion of an *environment*, which is a probability measure $\mathcal{E}$ on the set of tasks. Thus $\mathcal{E}(\tau)$ is

the probability of encountering the task τ in the environment $\mathcal{E}$, and $\mu_\tau(x, y)$ is the probability of finding the pair (x, y) in the context of the task τ.

Given $\mathcal{E}$ the *transfer risk* (or simply risk) of a learning algorithm A is defined as follows. We draw a task from the environment, $\tau \sim \mathcal{E}$, which fixes a corresponding distribution μ_τ on $H \times \mathbb{R}$. Then we draw a training sample $\mathbf{z} \sim \mu_\tau^m$ and use the algorithm to compute the predictor $A(\mathbf{z})$. Finally we measure the performance of this predictor on test points $(x, y) \sim \mu_\tau$. The corresponding definition of the transfer risk of A reads as

$$R_{\mathcal{E}}(A) = \mathbb{E}_{\tau\sim\mathcal{E}}\mathbb{E}_{\mathbf{z}\sim\mu_\tau^m}\mathbb{E}_{(x,y)\sim\mu_\tau}\left[\ell\left(\langle A(\mathbf{z}), x\rangle, y\right)\right],$$

which is simply the expected loss incurred by the use of the algorithm A on tasks drawn from the environment $\mathcal{E}$.

For any given dictionary $D \in \mathcal{D}_K$ we consider the learning algorithm A_D, which for $\mathbf{z} \in \mathcal{Z}^m$ computes the predictor

$$A_D(\mathbf{z}) = D\ \arg\min_{\gamma\in\mathcal{C}_\alpha} \frac{1}{m}\sum_{i=1}^{m} \ell\left(\langle D\gamma, x_i\rangle, y_i\right).$$

Equivalently, we can regard A_D as the Lasso operating on data preprocessed by the linear map $D^\top$, the adjoint of D.

We can make a single observation of the environment $\mathcal{E}$ in the following way: one first draws a task $\tau \sim \mathcal{E}$. This task and the corresponding distribution μ_τ are then observed by drawing an i.i.d. sample $\mathbf{z}$ from μ_τ, that is $\mathbf{z} \sim \mu_\tau^m$. For simplicity the sample size m will be fixed. Such an observation corresponds to the draw of a sample $\mathbf{z}$ from a probability distribution $\rho_{\mathcal{E}}$ on $(H \times \mathbb{R})^m$ which is defined by

$$\rho_{\mathcal{E}}(\mathbf{z}) := \mathbb{E}_{\tau\sim\mathcal{E}}\left[(\mu_\tau)^m(\mathbf{z})\right].$$

To estimate an environment a large number T of independent observations is needed, corresponding to a vector $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_T) \in ((H \times \mathbb{R})^m)^T$ drawn i.i.d. from $\rho_{\mathcal{E}}$, that is $\mathbf{Z} \sim (\rho_{\mathcal{E}})^T$.

We now propose to solve the problem (2.1) with the data $\mathbf{Z}$, ignore the resulting $\gamma_i(\mathbf{Z})$, but retain the dictionary $D(\mathbf{Z})$ and use the algorithm $A_{D(\mathbf{Z})}$ on future tasks drawn from the same environment. The performance of this method can be quantified as the transfer risk $R_{\mathcal{E}}\left(A_{D(\mathbf{Z})}\right)$ as defined above in (3.2) and again we are interested in comparing this to the risk of an ideal solution based on complete knowledge of the environment. For any fixed dictionary D and task τ the best we can do is to choose $\gamma \in \mathcal{C}$ so as to minimize $\mathbb{E}_{(x,y)\sim\mu_\tau}\left[\ell\left(\langle D\gamma, x\rangle, y\right)\right]$, so the best is to choose D so as to minimize the average of this over $\tau \sim \mathcal{E}$. The quantity

$$R_{\mathrm{opt}} = \min_{D\in\mathcal{D}_K} \mathbb{E}_{\tau\sim\mathcal{E}} \min_{\gamma\in\mathcal{C}_\alpha} \mathbb{E}_{(x,y)\sim\mu_\tau} \ell\left[\left(\langle D\gamma, x\rangle, y\right)\right]$$

thus describes the optimal performance achievable under the given constraint. Our second result is

Theorem 2. *With probability at least $1-\delta$ in the multisample $\mathbf{Z} = (\mathbf{X}, \mathbf{Y}) \sim \rho_{\mathcal{E}}^T$ we have*

$$R_{\mathcal{E}}\left(A_{D(\mathbf{Z})}\right) - R_{\mathrm{opt}} \le L\alpha K\sqrt{\frac{2\pi S_1(\mathbf{X})}{T}} + 4L\alpha\sqrt{\frac{S_\infty(\mathcal{E})(2+\ln K)}{m}} + \sqrt{\frac{8\ln 4/\delta}{T}},$$

where $S_1(\mathbf{X})$ *is as in Theorem 1 and* $S_\infty(\mathcal{E}) := \mathbb{E}_{\tau\sim\mathcal{E}}\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mu_\tau^m}\lambda_{\max}\left(\hat{\Sigma}(\mathbf{x})\right)$.

We discuss some implications of the above theorem. Some of these are analogous to the remarks following Theorem 1.

1. The interpretation of $S_\infty(\mathcal{E})$ is analogous to that of $S_\infty(\mathbf{X})$ in the bound for Theorem 1. The same applies to Remark 6 following Theorem 1.

2. In the regime $T \leq K^2$ the result does not imply any useful behavior. On the other and, if $T \gg K^2$ the dominant term in the bound is of order $\sqrt{S_\infty(\mathcal{E})/m}$.

3. There is an important difference with the multitask learning bound, namely in Theorem 2 we have $\sqrt{T}$ in the denominator of the first term of the excess risk, and not $\sqrt{mT}$ as in Theorem 1. This is because in the setting of learning to learn there is always a possibility of being misled by the draw of the training tasks. This possibility can only decrease as T increases – increasing m does not help.

The proof of Theorem 2 is given in Section B.2 of the supplementary appendix and follows the method outlined in [22]: one first bounds the estimation error for the expected empirical risk on future tasks, and then combines this with a bound of the expected true risk by said expected empirical risk. The term $K/\sqrt{T}$ may be an artifact of our method of proof and the conjecture that it can be replaced by $\sqrt{K/T}$ seems plausible.

3.3 Connection to sparse coding

We discuss a special case of Theorem 2 in the limit $m \to \infty$, showing that it subsumes the sparse coding result in [23]. To this end, we assume the noiseless generative model $y_{ti} = \langle w_t, x_{ti}\rangle$ described in Section 2, that is $\mu(x,y) = p(x)\delta(y, \langle w, x\rangle)$, where p is the uniform distribution on the sphere in $\mathbb{R}^d$ (ie. the Haar measure). In this case the environment of tasks is fully specified by a measure ρ on the unit ball in $\mathbb{R}^d$ from which a task $w \in \mathbb{R}^d$ is drawn and the measure μ is identified with the vector w. Note that we do not assume that these tasks are obtained as sparse combinations of some dictionary. Under the above assumptions and choosing ℓ to be the square loss, we have that $\mathbb{E}_{(x,y)\sim\mu_t}\ell(\langle w, x\rangle, y) = \|w_t - w\|^2$. Consequently, in the limit of $m \to \infty$ method (2.1) reduces to a constrained version of sparse coding [25, 26], namely

$$\min_{D\in\mathcal{D}_K} \frac{1}{T}\sum_{t=1}^{T} \min_{\gamma\in\mathcal{C}_\alpha} \|D\gamma - w_t\|^2.$$

In turn, the transfer error of a dictionary D is given by the quantity $R(D) := \min_{\gamma\in\mathcal{C}_\alpha}\|D\gamma - w\|^2$ and $R_{\text{opt}} = \min_{D\in\mathcal{D}_K}\mathbb{E}_{w\sim\rho}\min_{\gamma\in\mathcal{C}_\alpha}\|D\gamma - w\|^2$. Given the constraints $D \in \mathcal{D}_K$, $\gamma \in \mathcal{C}_\alpha$ and $\|x\| \leq 1$, the square loss $\ell(y, y') = (y - y')^2$, evaluated at $y = \langle D\gamma, x\rangle$, can be restricted to the interval $y \in [-\alpha, \alpha]$, where it has the Lipschitz constant $2(1+\alpha)$ for any $y' \in [-1, 1]$, as is easily verified. Since $S_1(\mathbf{X}) = 1$ and $S_\infty(\mathcal{E}) < \infty$, the bound in Theorem 2 becomes

$$R(D) - R_{\text{opt}} \leq 2\alpha(1+\alpha)K\sqrt{\frac{2\pi}{T}} + 8\sqrt{\frac{\ln 4/\delta}{T}} \tag{3.1}$$

in the limit $m \to \infty$. The typical choice for α is $\alpha \leq 1$, which ensures that $\|D\gamma\| \leq 1$. In this case inequality (3.1) provides an improvement over the sparse coding bound in [23] (cf. Theorem 2 and Section 2.4 therein), which contains an additional term of the order of $\sqrt{(\ln T)/T}$ and the same leading term in K as in (3.1) but with slightly worse constant (14 instead of $4\sqrt{2\pi}$). The connection of our method to sparse coding is experimentally demonstrated in Section 4.3 and illustrated in Figure 5.

4 Experiments

In this section, we present experiments on a synthetic and a real datasets. The aim of the experiments is to study the statistical performance of the proposed method, in both settings of multitask learning and learning to learn. We compare our method, denoted as Sparse Coding Multi Task Learning (SC-MTL), with single task learning (independent ridge regression, RR) as a base line and multitask feature learning (MTFL) [3]. We also report on sensitivity analysis of the proposed method versus different number of parameters involved.

4.1 Optimization algorithm

We solve problem (2.1) by alternating minimization over the dictionary matrix D and the code vectors $\boldsymbol{\gamma}$. The techniques we use are very similar to standard methods for sparse coding and dictionary learning, see [15] and references therein for more information. Briefly, assuming that the loss function ℓ is convex and has Lipschitz continuous gradient, either minimization problem is convex and can be solved efficiently by proximal gradient methods, e.g. [8, 9]. The key ingredient in each step is the computation of the proximity operator, which in either problem has a closed form expression.

4.2 Toy experiment

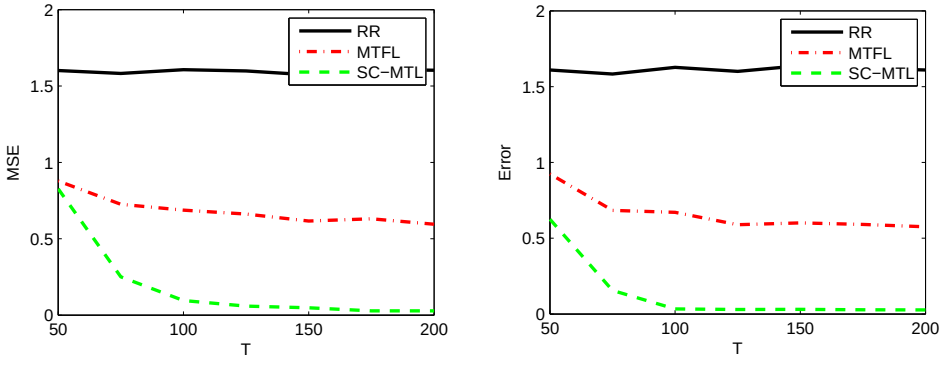

Figure 1: Multitask error (Left) and Transfer error (Right) vs. number of training tasks T.

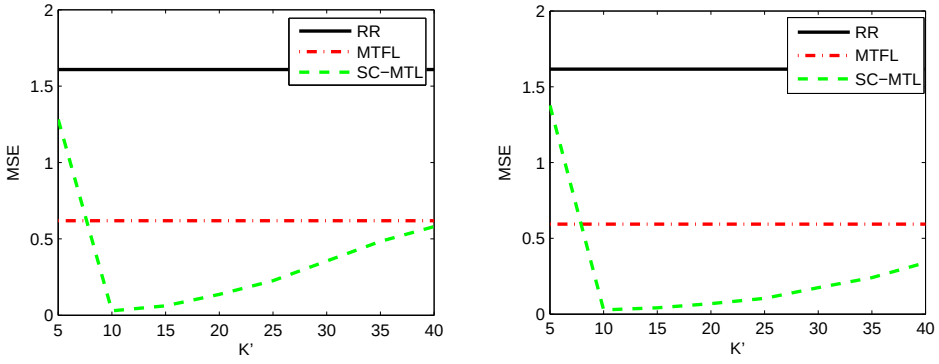

Figure 2: Multitask error (Left) and Transfer error (Right) vs. number of atoms K' of used by our method.

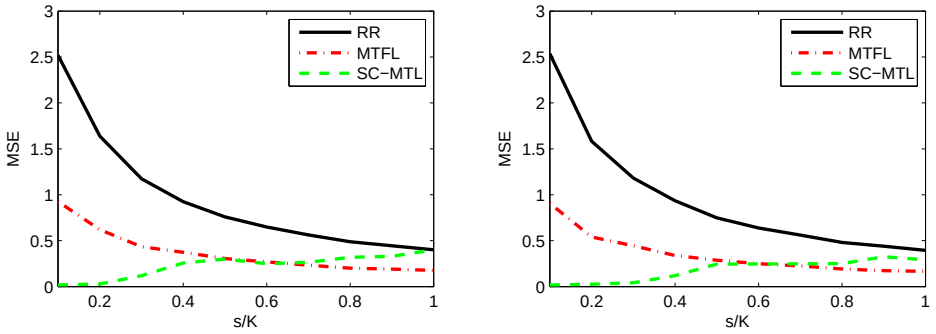

Figure 3: Multitask error (Left) and Transfer error (Right) vs. sparsity ratio s/K.

We generated a synthetic environment of tasks as follows. We choose a $d \times K$ matrix D by sampling its columns independently from the uniform distribution on the unit sphere in $\mathbb{R}^d$. Once D is created, a generic task in the environment is given by $w = D\gamma$, where γ is an s-sparse vector obtained as follows. First, we generate a set $J \subseteq \{1, \ldots, K\}$ of cardinality s, whose elements (indices) are sampled uniformly without replacement from the set $\{1, \ldots, K\}$. We then set $\gamma_j = 0$ if $j \notin J$ and otherwise sample $\gamma_j \sim \mathcal{N}(0, 0.1)$. Finally, we normalize γ so that it has ℓ_1-norm equal to some prescribed value α. Using the above procedure we generated T tasks $w_t = D\gamma_t$, $t = 1, \ldots, T$. Further, for each task t we generated a training set $\mathbf{z}_t = \{(x_{ti}, y_{ti})\}_{i=1}^m$, sampling x_{ti} i.i.d. from the uniform distribution on the unit sphere in $\mathbb{R}^d$. We then set $y_{ti} = \langle w_t, x_{ti} \rangle + \xi_{ti}$, with $\xi_{ti} \sim \mathcal{N}(0, \sigma^2)$, where σ is the variance of the noise. This procedure also defines the generation of new tasks in the transfer learning experiments below. We note that since the input distribution is uniform on a high dimensional sphere, neither sparse coding nor PCA will produce a useful representation, so there is no point in comparing to these methods here.

The above model depends on seven parameters: the number K and the dimension d of the atoms, the sparsity s and the ℓ_1-norm α of the codes, the noise level σ, the sample size per task

m and the number of training tasks T. In all experiments we report both the multitask learning (MTL) and learning to learn (LTL) performance of the methods studied. For MTL, we measure performance by the estimation error $1/T \sum_{t=1}^{T} \|w_t - \hat{w}_t\|^2$, where $\hat{w}_1, \ldots, \hat{w}_T$ are the estimated task vectors (in the case of our method $\hat{w}_t = D(\mathbb{Z})\gamma(\mathbb{Z})_t$ – see the discussion in Section 2. For LTL, we use the same quantity but with a new set of tasks generated by the environment (in the experiment below we generate 100 new tasks). The regularization parameter of each method is chosen by cross validation. Finally, all experiments are repeated 50 times, and the average performance results are reported in the plots below.

In the first experiment, we fix $K = 10, d = 20, s = 2, \alpha = 10, m = 10, \sigma = 0.1$ and study the statistical performance of the methods as a function of the number of tasks. The results, shown in Figure 1 clearly indicate that the proposed method outperforms both ridge regression and multitask feature learning. In this experiment the number of atoms used by our method, which here we denote by K' to avoid confusion with the number of atoms K of the target dictionary, was equal to $K = 10$, which gives an advantage to our method. We therefore also studied the performance of the method in dependence on K' . Figure 2, reporting this result, is in qualitative agreement with our theoretical analysis: the performance of the method is not too sensitive to K' if $K' \geq K$, and the method still outperforms independent task learning and multitask feature learning if $K' = 4K$. On the other hand if $K' < K$ the performance of the method quickly degrades. In the last experiment we study performance vs. the sparsity ratio s/K. Intuitively we would expect our method to have greater advantage over multitask feature learning if $s \ll K$. The results, shown in Figure 3, confirm this fact, also indicating that our method is outperformed by multitask feature learning method as sparsity becomes less pronounced ($s/K > 0.6$).

4.3 Sparse coding of images with missing pixels

In the next experiment we consider a sparse coding problem [25] of optical character images, with missing pixels. We employ the Binary Alphadigits dataset[1], which is composed of a set of binary 20×16 images of all digits and capital letters (39 images for each character). In the following experiment only the digits are used. We regard each image as a task, hence the input space is the set of 320 possible pixels indices, while the output space is the real interval $[0, 1]$, representing the gray level. We sample $T = 100, 130, 160, 190, 220, 250$ images, equally divided among the 10 possible digits. For each of these, a corresponding random set of $m = 160$ pixel values are sampled (so the set of sample pixels varies from one image to another).

We test the performance of the dictionary learned by method (2.1) in a learning to learn setting, by choosing 100 new images. The regularization parameter for each approach is tuned using cross validation. The results (Figure 4) indicate some advantage of the proposed method over trace norm regularization. For a more thorough understanding of the results, let us recall that MTFL assumes that there is a common representation of all data across the tasks. Therefore, the lack of a big improvement of MTFL is probably due to the fact that there are 10 different groups of tasks (corresponding to the 10 digits) so that tasks in different groups need a different representation of the data. This is an instance of negative transfer, already mentioned

[1] Available at *http://www.cs.nyu.edu/∼roweis/data.html*.

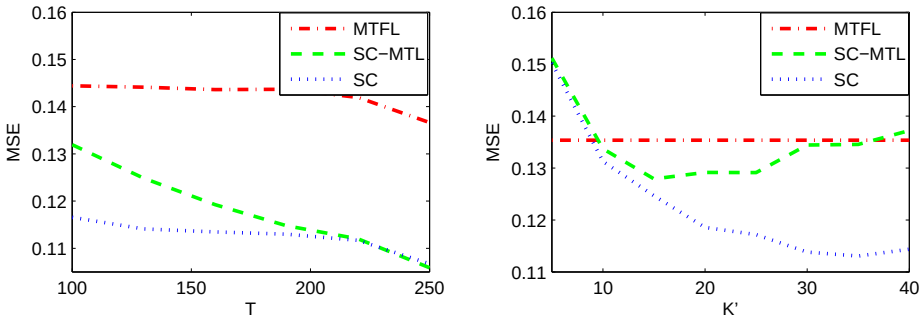

Figure 4: Transfer error vs. number of tasks T (Left) and vs. number of atoms K (Right) on the Binary Alphadigits dataset.

in Section 3.1. In contrast, SC-MTL assumes that each task will use only a small subset of the learned dictionary elements, thereby overcoming this limitation. A similar trend, not reported here due to space constraints, is obtained in the multitask setting. Ridge regression performed significantly worse and is not shown in the figure. We also show as a reference the performance of sparse coding (SC) applied to the T complete images, each of which corresponds to a task in the MTL formulation. Thus SC can be seen as applying SC-MTL for image completion, when all pixels are known.

With the aim of analyzing the atoms learned by the algorithm, we have carried out another experiment where we assume that there are 10 underlying atoms (one for each digit). We compare the resultant dictionary to that obtained by sparse coding, where all pixels are known. The results are shown in Figure 5 and the similarity of the dictionaries confirms the theoretical findings of Section 3.3.

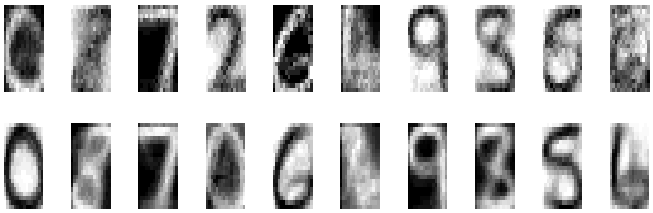

Figure 5: Dictionaries found by SC-MTL using $m = 240$ pixels (missing 25% pixels) per image (top) and by Sparse Coding employing all pixels (bottom).

5 Summary

In this paper, we have explored an application of sparse coding, which has been widely used in unsupervised learning and signal processing, to the domains of multitask learning and learning to learn. Our learning bounds provide a justification of this method and offer insights into

its advantage over independent task learning and learning dense representation of the tasks. The bounds, which hold in a Hilbert space setting, depend on data dependent quantities which measure the intrinsic dimensionality of the data. Numerical simulations presented here, as well as recent empirical results in [17] indicate that sparse coding is a promising approach to multitask learning and can lead to significant improvements over competing methods.

In the future, it would be valuable to study extensions of our analysis to more general classes of code vectors. For example, we could use code sets $\mathcal{C}_\alpha$ which arise from structured sparsity norms, such as the group Lasso norm, non overlapping groups [15] or other families of regularizers. A concrete example which comes to mind is to choose $K = Qr$, $Q, r \in \mathbb{N}$ and a partition $\mathcal{J} = \{\{(q-1)r+1, \ldots, qr\} : q = 1, \ldots, Q\}$ of the index set $\{1, \ldots, K\}$ into contiguous index sets of size r. Then using a norm of the type $\|\gamma\| = \|\gamma\|_1 + \sum_{J \in \mathcal{J}} \|\gamma_J\|_2$ will encourage codes which are sparse and use only few of the groups in $\mathcal{J}$. Using the ball associated with this norm as our set of codes would allow to model sets of tasks which are divided into groups.

References

[1] R.K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *J. Machine Learning Research*, 6:1817–1853, 2005.

[2] M. Anthony and P. Bartlett. *Learning in Neural Networks: Theoretical Foundations*, Cambridge University Press, 1999.

[3] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[4] A. Argyriou, A. Maurer, and M. Pontil. An algorithm for transfer learning in a heterogeneous environment. In Proceedings of ECML/PKDD (1), pages 71–85, 2008.

[5] B. Bakker and T. Heskes. Task clustering and gating for Bayesian multi–task learning. *J. Machine Learning Research*, 4:83–99, 2003.

[6] P.L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: risk bounds and structural results. *J. Machine Learning Research*, 3: 463–482, 2002.

[7] J. Baxter. A model for inductive bias learning. *J. of Artificial Intelligence Research*, 12:149–198, 2000.

[8] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal of Imaging Sciences*, 2(1):183–202, 2009.

[9] P.L. Combettes and V.R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling and Simulation*, 4(4):1168–1200, 2006.

[10] S. Ben-David and R. Schuller. Exploiting task relatedness for multiple task learning. Proceedings of Computational Learning Theory (COLT), 2003.

[11] P.J. Bickel, Y. Ritov, A.B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector *The Annals of Statistics*, 37(4):1705–1732, 2009.

[12] P. Bühlmann, S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, 2011.

[13] R. Caruana. Multi-task learning. *Machine Learning*, 28:41–75, 1997.

[14] T. Evgeniou, C.A. Micchelli, M. Pontil. Learning multiple tasks with kernel methods. *J. Machine Learning Research*, 6:615–637, 2005.

[15] R. Jenatton, J. Mairal, G. Obozinski, F. Bach. Proximal methods for hierarchical sparse coding. *J. Machine Learning Research*, 12:2297–2334, 2011

[16] V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 30(1):1–50, 2002.

[17] A. Kumar, H. Daumé III. Learning task grouping and overlap in multi-task learning. International Conference on Machine Learning (ICML), 2012

[18] M. Ledoux and M. Talagrand. *Probability in Banach Spaces*, Springer 1991.

[19] J. Mairal, F. Bach, J. Ponce. Task-driven dictionary learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):791–804, 2012.

[20] A. Maurer. Bounds for linear multi-task learning. *Journal of Machine Learning Research*, 7:117–139, 2006.

[21] A. Maurer. Concentration inequalities for functions of independent variables. *Random Structures and Algorithms*, 29:121–138, 2006.

[22] A. Maurer. Transfer bounds for linear feature learning. *Machine Learning*, 75(3):327–350, 2009.

[23] A. Maurer and M. Pontil. K-dimensional coding schemes in Hilbert spaces. *IEEE Transactions on Information Theory*, 56(11):5839–5846, 2010.

[24] C. McDiarmid. *Concentration*, in *Probabilistic Methods of Algorithmic Discrete Mathematics*, pages 195–248, Springer, 1998.

[25] B.A. Olshausen. Learning linear, sparse, factorial codes. A.I. Memo 1580, Massachusetts Institute of Technology, 1996.

[26] B.A. Olshausen and D.J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[27] R. Raina, A. Battle, H. Lee, B. Packer and A.Y. Ng Self-taught Learning: Transfer Learning from Unlabeled Data. Proceedings of the Twenty-Fourth International Conference on Machine Learning, 2007, AAAI Press.

[28] D. Slepian. The one-sided barrier problem for Gaussian noise. *Bell System Tech. J.*, 41:463–501, 1962.

[29] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B*, 58(1):267–288, 1996.

[30] S. Thrun and L. Pratt. *Learning to Learn*. Springer, 1998.

[31] S.A. van de Geer. High-dimensional generalized linear models and the Lasso. *Annals of Statistics*, 36(2):614-645, 2008.

Appendix

In this appendix, we present the proof of Theorems 1 and 2. We begin by introducing some more notation and auxiliary results.

A Notation and tools

Issues of measurability will be ignored throughout, in particular, if $\mathcal{F}$ is a class of real valued functions on a domain $\mathcal{X}$ and X a random variable with values in $\mathcal{X}$ then we will always write $\mathbb{E}\sup_{f\in\mathcal{F}} f(X)$ to mean $\sup\{\mathbb{E}\max_{f\in\mathcal{F}_0} f(X) : \mathcal{F}_0 \subseteq \mathcal{F}, \mathcal{F}_0 \text{ finite}\}$.

In the sequel H denotes a finite or infinite dimensional Hilbert space with inner product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\|$. If T is a bounded linear operator on H its operator norm is written $\|T\|_\infty = \sup\{\|Tx\| : \|x\| = 1\}$.

Members of H are denoted with lower case italics such as x, v, w, vectors composed of such vectors are in bold lower case, i.e. $\mathbf{x} = (x_1, \ldots, x_m)$ or $\mathbf{v} = (v_1, \ldots, v_n)$, where m or n are explained in the context.

An *example* is a pair $z = (x, y) \in B \times \mathbb{R} =: \mathcal{Z}$, a sample is a vector of such pairs $\mathbf{z} = (z_1, \ldots, z_m) = ((x_1, y_1), \ldots, (x_m, y_m))$. Here we also write $\mathbf{z} = (\mathbf{x}, \mathbf{y})$, with $\mathbf{x} = (x_1, \ldots, x_m) \in H^m$ and $\mathbf{y} = (y_1, \ldots, y_m) \in \mathbb{R}^m$.

A multisample is a vector $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_T)$ composed of samples. We also write $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$ with $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$.

For members of $\mathbb{R}^K$ we use the greek letters γ or β. Depending on context the inner product and euclidean norm on $\mathbb{R}^K$ will also be denoted with $\langle\cdot,\cdot\rangle$ and $\|.\|$. The ℓ_1-norm $\|\cdot\|_1$ on $\mathbb{R}^K$ is defined by $\|\beta\|_1 = \sum_{k=1}^K |\gamma_k|$.

In the sequel we denote with $\mathcal{C}_\alpha$ the set $\left\{\beta \in \mathbb{R}^K : \|\beta\|_1 \leq \alpha\right\}$, abbreviate $\mathcal{C}$ for the ℓ_1-unit ball $\mathcal{C}_1$. The canonical basis of $\mathbb{R}^K$ is denoted $e_1, \ldots, e_K$. Unless otherwise specified the summation over he index i will always run from 1 to m, t will run from 1 to T, and k will run from 1 to K.

A.1 Covariances

For $\mathbf{x} \in H^m$ the empirical covariance operator $\hat{\Sigma}(\mathbf{x})$ is specified by

$$\left\langle \hat{\Sigma}(\mathbf{x}) v, w \right\rangle = \frac{1}{m} \sum_i \langle v, x_i \rangle \langle x_i, w \rangle, \ v, w \in H.$$

The definition implies the inequality

$$\sum_i \langle v, x_i \rangle^2 = m \left\langle \hat{\Sigma}(\mathbf{x}) v, v \right\rangle \leq m \left\| \hat{\Sigma}(\mathbf{x}) \right\|_\infty \|v\|^2.$$

It also follows that $\operatorname{tr}\left(\hat{\Sigma}(\mathbf{x})\right) = (1/m) \sum_i \|x_i\|^2$.

For a multisample $\mathbf{X} \in H^{mT}$ we will consider two quantities defined in terms of the empirical covariances.

$$\begin{aligned} S_1(\mathbf{X}) &= \frac{1}{T} \sum_t \left\| \hat{\Sigma}(\mathbf{x}_t) \right\|_1 := \frac{1}{T} \sum_t \operatorname{tr}\left(\hat{\Sigma}(\mathbf{x}_t)\right) \\ S_\infty(\mathbf{X}) &= \frac{1}{T} \sum_t \left\| \hat{\Sigma}(\mathbf{x}_t) \right\|_\infty := \frac{1}{T} \sum_t \lambda_{\max}\left(\hat{\Sigma}(\mathbf{x}_t)\right), \end{aligned}$$

where $\lambda_{\max}$ is the largest eigenvalue. If all data points x_{ti} lie in the unit ball of H then $S_1(\mathbf{X}) \leq 1$. Of course $S_1(\mathbf{X})$ can also be written as the trace of the total covariance $(1/T) \sum_t \hat{\Sigma}(\mathbf{x}_t)$, while $S_\infty(\mathbf{X})$ will always be at least as large as the largest eigenvalue of the total covariance. We always have $S_\infty(\mathbf{X}) \leq S_1(\mathbf{X})$, with equality only if the data is one-dimensional for all tasks. The quotient $S_1(\mathbf{X}) / S_\infty(\mathbf{X})$ can be regarded as a crude measure of the effective dimensionality of the data. If the data have a high dimensional distribution for each task then $S_\infty(\mathbf{X})$ can be considerably smaller than $S_1(\mathbf{X})$.

A.2 Concentration inequalities

Let $\mathcal{X}$ be any space. For $\mathbf{x} \in \mathcal{X}^n$, $1 \leq k \leq n$ and $y \in \mathcal{X}$ we use $\mathbf{x}_{k \leftarrow y}$ to denote the object obtained from $\mathbf{x}$ by replacing the k-th coordinate of $\mathbf{x}$ with y. That is

$$\mathbf{x}_{k \leftarrow y} = (x_1, \ldots, x_{k-1}, y, x_{k+1}, \ldots, x_n).$$

The concentration inequality in part (i) of the following theorem, known as the bounded difference inequality is given in [24]. A proof of inequality (ii) is given in [21].

Theorem 3. *Let $F : \mathcal{X}^n \to \mathbb{R}$ and define A and B by*

$$\begin{aligned} A^2 &= \sup_{\mathbf{x} \in \mathcal{X}^n} \sum_{k=1}^n \sup_{y_1, y_2 \in \mathcal{X}} \left(F(\mathbf{x}_{k \leftarrow y_1}) - F(\mathbf{x}_{k \leftarrow y_2}) \right)^2 \\ B^2 &= \sup_{\mathbf{x} \in \mathcal{X}^n} \sum_{k=1}^n \left(F(\mathbf{x}) - \inf_{y \in \mathcal{X}} F(\mathbf{x}_{k \leftarrow y}) \right)^2. \end{aligned}$$

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a vector of independent random variables with values in $\mathcal{X}$, and let $\mathbf{X}'$ be i.i.d. to $\mathbf{X}$. Then for any $s > 0$

(i) $\Pr\{F(\mathbf{X}) > \mathbb{E}F(\mathbf{X}') + s\} \leq e^{-2s^2/A^2}$.

(ii) $\Pr\{F(\mathbf{X}) > \mathbb{E}F(\mathbf{X}') + s\} \leq e^{-s^2/\left(2B^2\right)}$.

A.3 Rademacher and Gaussian averages

We will use the term *Rademacher variables* for any set of independent random variables, uniformly distributed on $\{-1, 1\}$, and reserve the symbol σ for Rademacher variables. A set of random variables is called *orthogaussian* if the members are independent $\mathcal{N}(0, 1)$-distributed (standard normal) variables and reserve the letter ζ for standard normal variables. The notation $\sigma_1, \sigma_2, \ldots, \sigma_i, \ldots, \sigma_{11}, \ldots, \sigma_{ij}$ etc. will always refer to independent Rademacher variables and $\zeta_1, \zeta_2, \ldots, \zeta_i, \ldots, \zeta_{11}, \ldots, \zeta_{ij}$ will refer to orthogaussian variables.

For $A \subseteq \mathbb{R}^n$ we define the Rademacher and Gaussian averages of A ([18],[6]) as

$$\begin{aligned} \mathcal{R}(A) &= \mathbb{E}_\sigma \sup_{(x_1,\ldots,x_n)\in A} \frac{2}{n}\sum_{i=1}^n \sigma_i x_i, \\ \mathcal{G}(A) &= \mathbb{E}_\zeta \sup_{(x_1,\ldots,x_n)\in A} \frac{2}{n}\sum_{i=1}^n \zeta_i x_i. \end{aligned}$$

If $\mathcal{F}$ is a class of real valued functions on a space $\mathcal{X}$ and $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$ we write

$$\mathcal{F}(\mathbf{x}) = \mathcal{F}(x_1, \ldots, x_n) = \{(f(x_1), \ldots, f(x_n)) : f \in \mathcal{F}\} \subseteq \mathbb{R}^n.$$

The empirical Rademacher and Gaussian complexities of $\mathcal{F}$ on $\mathbf{x}$ are respectively $\mathcal{R}(\mathcal{F}(\mathbf{x}))$ and $\mathcal{G}(\mathcal{F}(\mathbf{x}))$.

The utility of these concepts for learning theory comes from the following key-result (see [6, 16]), stated here in two portions for convenience in the sequel.

Theorem 4. *Let $\mathcal{F}$ be a real-valued function class on a space $\mathcal{X}$ and $\mu_1, \ldots, \mu_m$ be probability measures on $\mathcal{X}$ with product measure $\boldsymbol{\mu} = \prod_i \mu_i$ on $\mathcal{X}^m$. For $\mathbf{x} \in \mathcal{X}^m$ define*

$$\Phi(\mathbf{x}) = \sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m \left(\mathbb{E}_{x\sim\mu_i}[f(x)] - f(x_i)\right).$$

Then $\mathbb{E}_{\mathbf{x}\sim\boldsymbol{\mu}}[\Phi(\mathbf{x})] \leq \mathbb{E}_{\mathbf{x}\sim\boldsymbol{\mu}}\mathcal{R}(\mathcal{F}(\mathbf{x}))$.

Proof. For any realization $\sigma = \sigma_1, \ldots, \sigma_m$ of the Rademacher variables

$$\begin{aligned} \mathbb{E}_{\mathbf{x}\sim\boldsymbol{\mu}}[\Phi(\mathbf{x})] &= \mathbb{E}_{\mathbf{x}\sim\boldsymbol{\mu}} \sup_{f\in\mathcal{F}} \frac{1}{m}\mathbb{E}_{\mathbf{x}'\sim\boldsymbol{\mu}} \sum_{i=1}^m (f(x_i') - f(x_i)) \\ &\leq \mathbb{E}_{\mathbf{x},\mathbf{x}'\sim\boldsymbol{\mu}\times\boldsymbol{\mu}} \sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m \sigma_i (f(x_i') - f(x_i)), \end{aligned}$$

because of the symmetry of the measure $\boldsymbol{\mu}\times\boldsymbol{\mu}(\mathbf{x}, \mathbf{x}') = \prod_i \mu_i \times \prod_i \mu_i(\mathbf{x}, \mathbf{x}')$under the interchange $x_i \leftrightarrow x_i'$. Taking the expectation in σ and applying the triangle inequality gives the result. ∎

Theorem 5. *Let $\mathcal{F}$ be a $[0,1]$-valued function class on a space $\mathcal{X}$, and $\boldsymbol{\mu}$ as above. For $\delta > 0$ we have with probability greater than $1-\delta$ in the sample $\mathbf{x} \sim \boldsymbol{\mu}$ that for all $f \in \mathcal{F}$*

$$\mathbb{E}_{x\sim\boldsymbol{\mu}}\left[f\left(x\right)\right] \leq \frac{1}{m}\sum_{i=1}^{m} f\left(x_i\right) + \mathbb{E}_{\mathbf{x}\sim\boldsymbol{\mu}}\mathcal{R}\left(\mathcal{F}\left(\mathbf{x}\right)\right) + \sqrt{\frac{\ln\left(1/\delta\right)}{2m}}.$$

To prove this apply the bounded-difference inequality (part (i) of Theorem 3) to the function Φ of the previous theorem (see e.g. [6]). Under the conditions of this result, changing one of the x_i will not change $\mathcal{R}\left(\mathcal{F}\left(\mathbf{x}\right)\right)$ by more than 2, so again by the bounded difference inequality applied to $\mathcal{R}\left(\mathcal{F}\left(\mathbf{x}\right)\right)$ and a union bound we obtain the data dependent version

Corollary 6. *Let $\mathcal{F}$ and $\boldsymbol{\mu}$ be as above. For $\delta > 0$ we have with probability greater than $1-\delta$ in the sample $\mathbf{x} \sim \boldsymbol{\mu}$ that for all $f \in \mathcal{F}$*

$$\mathbb{E}_{x\sim\boldsymbol{\mu}}\left[f\left(x\right)\right] \leq \frac{1}{m}\sum_{i=1}^{m} f\left(x_i\right) + \mathcal{R}\left(\mathcal{F}\left(\mathbf{x}\right)\right) + \sqrt{\frac{9\ln\left(2/\delta\right)}{2m}}.$$

To bound Rademacher averages the following result is very useful [6, 1, 18]

Lemma 7. *Let $A \subseteq \mathbb{R}^n$, and let $\psi_1, \ldots, \psi_n$ be real functions such that $\psi_i\left(s\right) - \psi_i\left(t\right) \leq L\left|s-t\right|$,$\forall i$, and $s, t \in \mathbb{R}$. Define $\boldsymbol{\psi}\left(A\right) = \left\{\psi_1\left(x_1\right), \ldots, \psi_n\left(x_n\right) : \left(x_1, \ldots, x_n\right) \in A\right\}$. Then*

$$\mathcal{R}\left(\boldsymbol{\psi}\left(A\right)\right) \leq L\mathcal{R}\left(A\right).$$

Sometimes it is more convenient to work with Gaussian averages which can be used instead, by virtue of the next lemma. For a proof see, for example, [18, p. 97]

Lemma 8. *For $A \subseteq \mathbb{R}^k$ we have $\mathcal{R}\left(A\right) \leq \sqrt{\pi/2}\,\mathcal{G}\left(A\right)$.*

The next result is known as Slepian's lemma ([28], [18]).

Theorem 9. *Let Ω and Ξ be mean zero, separable Gaussian processes indexed by a common set $\mathcal{S}$, such that*

$$\mathbb{E}\left(\Omega_{s_1} - \Omega_{s_2}\right)^2 \leq \mathbb{E}\left(\Xi_{s_1} - \Xi_{s_2}\right)^2 \text{ for all } s_1, s_2 \in \mathcal{S}.$$

Then

$$\mathbb{E}\sup_{s\in\mathcal{S}} \Omega_s \leq \mathbb{E}\sup_{s\in\mathcal{S}} \Xi_s.$$

B Proofs

B.1 Multitask learning

In this section we prove Theorem 1. It is an immediate consequence of Hoeffding's inequality and the following uniform bound on the estimation error.

Theorem 10. *Let $\delta > 0$, fix K and let $\mu_1, \ldots, \mu_T$ be probability measures on $H \times \mathbb{R}$. With probability at least $1 - \delta$ in the draw of $\mathbf{Z} \sim \prod_{t=1}^{T} (\mu_t)$ we have for all $D \in \mathcal{D}_K$ and all $\boldsymbol{\gamma} \in \mathcal{C}_\alpha^T$ that*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{(x,y)\sim\mu_t} \left[\ell\left(\left\langle D\gamma_t, x\right\rangle, y\right)\right] - \frac{1}{mT}\sum_{t=1}^{T}\sum_{i=1}^{m} \ell\left(\left\langle D\gamma_t, x_{ti}\right\rangle, y_{ti}\right)$$
$$\leq L\alpha\sqrt{\frac{2S_1\left(\mathbf{X}\right)\left(K+12\right)}{mT}} + L\alpha\sqrt{\frac{8S_\infty\left(\mathbf{X}\right)\ln\left(2K\right)}{m}} + \sqrt{\frac{9\ln 2/\delta}{2mT}}.$$

The proof of this theorem requires auxiliary results. Fix $\mathbf{X} \in H^{mT}$ and for $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_T) \in \left(\mathbb{R}^K\right)^T$ define the random variable

$$F_{\boldsymbol{\gamma}} = F_{\boldsymbol{\gamma}}\left(\boldsymbol{\sigma}\right) = \sup_{D\in\mathcal{D}_K} \sum_{t,i} \sigma_{ti}\left\langle D\gamma_t, x_{ti}\right\rangle.$$

Lemma 11. *(i) If $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_T)$ satisfies $\|\gamma_t\| \leq 1$ for all t, then*

$$\mathbb{E}F_{\boldsymbol{\gamma}} \leq \sqrt{mTK\ S_1\left(\mathbf{X}\right)}.$$

(ii) If $\boldsymbol{\gamma}$ satisfies $\|\gamma_t\|_1 \leq 1$ for all t, then for any $s \geq 0$

$$\Pr\left\{F_{\boldsymbol{\gamma}} \geq \mathbb{E}\left[F_{\boldsymbol{\gamma}}\right] + s\right\} \leq \exp\left(\frac{-s^2}{8mT\ S_\infty\left(\mathbf{X}\right)}\right).$$

Proof. (i) We observe that

$$\begin{aligned}
\mathbb{E}F_{\boldsymbol{\gamma}} &= \mathbb{E}\sup_D \sum_k \left\langle De_k, \sum_{t,i} \sigma_{ti}\gamma_{tk}x_{ti}\right\rangle \\
&\leq \sup_D \left(\sum_k \|De_k\|^2\right)^{1/2} \mathbb{E}\left(\sum_k \left\|\sum_{t,i} \sigma_{ti}\gamma_{tk}x_{ti}\right\|^2\right)^{1/2} \\
&\leq \sqrt{K}\left(\sum_k |\gamma|_{tk}^2\, \mathbb{E}\left\|\sum_{t,i} \sigma_{ti}x_{ti}\right\|^2\right)^{1/2} \leq \sqrt{K\sum_{t,i}\|x_{ti}\|^2} = \sqrt{mTK\ S_1\left(\mathbf{X}\right)}.
\end{aligned}$$

(ii) For any configuration $\boldsymbol{\sigma}$ of the Rademacher variables let $D\left(\boldsymbol{\sigma}\right)$ be the maximizer in the definition of $F_{\boldsymbol{\gamma}}\left(\boldsymbol{\sigma}\right)$. Then for any $s \in \{1, \ldots, T\}$, $j \in \{1, \ldots, m\}$ and any $\sigma' \in \{-1, 1\}$ to replace σ_{sj} we have

$$F_{\boldsymbol{\gamma}}\left(\boldsymbol{\sigma}\right) - F_{\boldsymbol{\gamma}}\left(\boldsymbol{\sigma}_{(sj)\leftarrow\sigma'}\right) \leq 2\left|\left\langle D\left(\boldsymbol{\sigma}\right)\gamma_s, x_{sj}\right\rangle\right|.$$

Using the inequality (A.1) we then obtain

$$\begin{aligned}\sum_{sj}\left(F_{\gamma}\left(\boldsymbol{\sigma}\right)-\inf_{\sigma'\in\{-1,1\}}F_{\boldsymbol{\gamma}}\left(\boldsymbol{\sigma}_{(sj)\leftarrow\sigma'}\right)\right)^{2} &\leq 4\sum_{t,i}\left\langle D\left(\boldsymbol{\sigma}\right)\gamma_{t},x_{ti}\right\rangle^{2}\\ &\leq 4m\sum_{t}\left\|\hat{\Sigma}\left(\mathbf{x}_{t}\right)\right\|_{\infty}\left\|D\left(\boldsymbol{\sigma}\right)\gamma_{t}\right\|^{2}\\ &\leq 4m\sum_{t}\left\|\hat{\Sigma}\left(\mathbf{x}_{t}\right)\right\|_{\infty}.\end{aligned}$$

In the last inequality we used the fact that for any $D\in\mathcal{D}_{K}$ we have $\left\|D\gamma_{t}\right\|\leq\sum_{k}\left|\gamma_{tk}\right|\left\|De_{k}\right\|\leq\left\|\gamma_{t}\right\|_{1}\leq 1$. The conclusion now follows from part (ii) of Theorem 3. ■

Proposition 12. *For every fixed* $\mathbf{Z}=(\mathbf{X},\mathbf{Y})\in(H\times\mathbb{R})^{mT}$ *we have*

$$\mathbb{E}_{\sigma}\sup_{D\in D,\boldsymbol{\gamma}\in\left(\mathcal{C}_{\alpha}\right)^{T}}\sum_{t,i}\sigma_{it}\ell\left(\left\langle D\gamma_{t},x_{ti}\right\rangle,y_{ti}\right)\leq L\alpha\sqrt{2mTS_{1}\left(\mathbf{X}\right)\left(K+12\right)}+L\alpha T\sqrt{8mS_{\infty}\left(\mathbf{X}\right)\ln\left(2K\right)}.$$

Proof. It suffices to prove the result for $\alpha=1$, the general result being a consequence of rescaling. By Lemma 7 and the Lipschitz properties of the loss function ℓ we have

$$\mathbb{E}_{\sigma}\sup_{D\in\mathcal{D}_{K},\boldsymbol{\gamma}\in\left(\mathcal{C}\right)^{T},}\sum_{t,i}\sigma_{it}\ell\left(\left\langle D\gamma_{t},x_{ti}\right\rangle,y_{ti}\right)\leq L\mathbb{E}_{\sigma}\sup_{D\in\mathcal{D}_{K},\boldsymbol{\gamma}\in\left(\mathcal{C}\right)^{T},}\sum_{t,i}\sigma_{it}\left\langle D\gamma_{t},x_{ti}\right\rangle.$$

Since linear functions on a compact convex set attain their maxima at the extreme points, we have

$$\mathbb{E}\sup_{D\in\mathcal{D}_{K},\boldsymbol{\gamma}\in\left(\mathcal{C}\right)^{T},}\sum_{t=1}^{T}\sum_{i=1}^{m}\sigma_{it}\left\langle D\gamma_{t},x_{ti}\right\rangle=\mathbb{E}\max_{\boldsymbol{\gamma}\in\text{ext}\left(\mathcal{C}\right)^{T}}F_{\boldsymbol{\gamma}},$$

where $F_{\boldsymbol{\gamma}}$ is defined as in (B.1). Now for any $\delta\geq 0$ we have, since $F_{\boldsymbol{\gamma}}\geq 0$,

$$\begin{aligned}\mathbb{E}\max_{\boldsymbol{\gamma}\in\text{ext}\left(\mathcal{C}\right)^{T}}F_{\boldsymbol{\gamma}} &= \int_{0}^{\infty}\Pr\left\{\max_{\boldsymbol{\gamma}\in\text{ext}\left(\mathcal{C}\right)^{T}}F_{\boldsymbol{\gamma}}>s\right\}ds\\ &\leq \sqrt{mKTS_{1}\left(\mathbf{X}\right)}+\delta+\sum_{\boldsymbol{\gamma}\in\left(\text{ext}\left(\mathcal{C}\right)\right)^{T}}\int_{\sqrt{mKTS_{1}\left(\mathbf{X}\right)}+\delta}^{\infty}\Pr\left\{F_{\boldsymbol{\gamma}}>s\right\}ds\\ &\leq \sqrt{mKTS_{1}\left(\mathbf{X}\right)}+\delta+\sum_{\boldsymbol{\gamma}\in\left(\text{ext}\left(\mathcal{C}\right)\right)^{T}}\int_{\delta}^{\infty}\Pr\left\{F_{\boldsymbol{\gamma}}>\mathbb{E}F_{\boldsymbol{\gamma}}+s\right\}ds\\ &\leq \sqrt{mKTS_{1}\left(\mathbf{X}\right)}+\delta+\left(2K\right)^{T}\int_{\delta}^{\infty}\exp\left(\frac{-s^{2}}{8mTS_{\infty}\left(\mathbf{X}\right)}\right)ds\\ &\leq \sqrt{mKTS_{1}\left(\mathbf{X}\right)}+\delta+\frac{4mTS_{\infty}\left(\mathbf{X}\right)\left(2K\right)^{T}}{\delta}\exp\left(\frac{-\delta^{2}}{8mTS_{\infty}\left(\mathbf{X}\right)}\right).\end{aligned}$$

Here the first inequality follows from the fact that probabilities never exceed 1 and a union bound. The second inequality follows from Lemma 11, part (i), since $\mathbb{E}F_{\mathbf{k}}\leq\sqrt{mKTS_{1}\left(\mathbf{X}\right)}$.

The third inequality follows from Lemma 11, part (ii), and the fact that the cardinality of ext$(\mathcal{C})$ is $2K$, and the last inequality follows from a well known estimate on Gaussian random variables. Setting $\delta = \sqrt{8mTS_\infty(\mathbf{X}) \ln\left(e(2K)^T\right)}$ we obtain with some easy simplifying estimates

$$\mathbb{E} \max_{\boldsymbol{\gamma} \in \text{ext}(\mathcal{C})^T} F_{\boldsymbol{\gamma}} \leq \sqrt{2mT(K+12) S_1(\mathbf{X})} + T\sqrt{8mS_\infty(\mathbf{X}) \ln(2K)},$$

which together with (B.1) and (B.1) gives the result. ■

Theorem 10 now follows from Corollary 6.

If the set $\mathcal{C}_\alpha$ is replaced by any other subset $\mathcal{C}'$ of the ℓ_2-ball of radius α, a similar proof strategy can be employed. The denominator in the exponent of Lemma 11-(ii) then obtains another factor of $\sqrt{K}$. The union bound over the extreme points in ext$(\mathcal{C})$ in the previous proposition can be replaced by a union bound over a cover $\mathcal{C}'$. This leads to the alternative result mentioned in Remark 5 following the statement of Theorem 1.

Another modification leads to a bound for the method presented in [17], where the constraint $\|De_k\| \leq 1$ is replaced by $\|D\|_2 \leq \sqrt{K}$ (here $\|.\|_2$ is the Frobenius or Hilbert Schmidt norm) and the constraint $\|\gamma_t\|_1 \leq \alpha, \forall t$ is replaced by $\sum \|\gamma_t\|_1 \leq \alpha T$. To explain the modification we set $\alpha = 1$. Part (i) of Lemma 11 is easily verified. The union bound over $(\text{ext}(\mathcal{C}))^T$ in the previous proposition is replaced by a union bound over the $2TK$ extreme points of the ℓ_1-Ball of radius T in $\mathbb{R}^{TK}$. For part (ii) we use the fact that the concentration result is only needed for $\boldsymbol{\gamma}$ being an extreme point (so that it involves only a single task) and obtain the bound $\sum_t \left\|\hat{\Sigma}(\mathbf{x}_t)\right\|_\infty \|D\gamma_t\|^2 \leq TKS'_\infty(\mathbf{X})$, leading to

$$\Pr\{F_{\boldsymbol{\gamma}} \geq E[F_{\boldsymbol{\gamma}}] + s\} \leq \exp\left(\frac{-s^2}{8mTK\ S'_\infty(\mathbf{X})}\right).$$

Proceeding as above we obtain the excess risk bound

$$L\alpha\sqrt{\frac{2S_1(\mathbf{X})(K+12)}{mT}} + L\alpha\sqrt{\frac{8KS'_\infty(\mathbf{X}) \ln(2KT)}{m}} + \sqrt{\frac{8 \ln 4/\delta}{mT}},$$

to replace the bound in Theorem 1. The factor $\sqrt{K}$ in the second term seems quite weak, but it must be borne in mind that the constraint $\|D\|_2 \leq \sqrt{K}$ is much weaker than $\|De_k\| \leq 1$, and allows for a smaller approximation error. If we retain $\|De_k\| \leq 1$ and only modify the γ-constraint to $\sum \|\gamma_t\|_1 \leq \alpha T$ the $\sqrt{K}$ in the second term disappears and by comparison to Theorem 1 there is only and additional $\ln T$ and the switch from $S_\infty(\mathbf{X})$ to $S'_\infty(\mathbf{X})$, reflecting the fact that $\sum \|\gamma_t\|_1 \leq \alpha T$ is a much weaker constraint than $\|\gamma_t\|_1 \leq \alpha, \forall t$, so that, again, a smaller minimum in (2.1) is possible for the modified method.

B.2 Learning to learn

In this section we prove Theorem 2. The basic strategy is as follows. Recall the definition (3.2) of the measure $\rho_{\mathcal{E}}$, which governs the generation of a training sample in the environment $\mathcal{E}$. On

a given training sample $\mathbf{z} \sim \rho_{\mathcal{E}}$ the algorithm A_D as defined in (3.2) incurs the empirical risk

$$\hat{R}_D(\mathbf{z}) = \min_{\gamma \in \mathcal{C}_\alpha} \frac{1}{m} \sum_{i=1}^{m} \ell\left(\langle D\gamma, x_i \rangle, y_i\right).$$

The algorithm A_D, essentially being the Lasso, has very good estimation properties, so $\hat{R}_D(\mathbf{z})$ will be close to the true risk of A_D in the corresponding task. This means that we only really need to estimate the expected empirical risk $\mathbb{E}_{\mathbf{z}\sim\rho_{\mathcal{E}}} \hat{R}_D(\mathbf{z})$ of A_D on future tasks. On the other hand the minimization problem (2.1) can be written as

$$\min_{D \in \mathcal{D}_K} \frac{1}{T} \sum_{t=1}^{T} \hat{R}_D(\mathbf{z}_t) \text{ with } \mathbf{Z} = (z_1, \ldots, z_T) \sim (\rho_{\mathcal{E}})^T,$$

with dictionary $D(\mathbf{Z})$ being the minimizer. If $\mathcal{D}_K$ is not too large this should be similar to $\mathbb{E}_{\mathbf{z}\sim\rho_{\mathcal{E}}} \hat{R}_{D(\mathbf{Z})}(\mathbf{z})$. In the sequel we make this precise.

Lemma 13. *For $v \in H$ with $\|v\| \leq 1$ and $\mathbf{x} \in H^m$ let F be the random variable*

$$F = \left| \left\langle v, \sum_i \sigma_i x_i \right\rangle \right|.$$

Then (i) $\mathbb{E}F \leq \sqrt{m} \left\| \hat{\Sigma}(\mathbf{x}) \right\|_\infty^{1/2}$ and (ii) for $t \geq 0$

$$\Pr\{F > \mathbb{E}F + s\} \leq \exp\left(\frac{-s^2}{2m \left\| \hat{\Sigma}(\mathbf{x}) \right\|_\infty} \right).$$

Proof. (i). Using Jensen's inequality and (A.1) we get

$$\mathbb{E}F \leq \left(\mathbb{E} \left\langle v, \sum_i \sigma_i x_i \right\rangle^2 \right)^{1/2} = \left(\sum_i \langle v, x_i \rangle^2 \right)^{1/2} \leq m \left\| \hat{\Sigma}(\mathbf{x}) \right\|_\infty.$$

(ii) Let $\boldsymbol{\sigma}$ be any configuration of the Rademacher variables. For any $\sigma', \sigma'' \in \{-1, 1\}$ to replace σ_{sj} we have

$$F\left(\boldsymbol{\sigma}_{(sj)\leftarrow\sigma'}\right) - F\left(\boldsymbol{\sigma}_{(sj)\leftarrow\sigma''}\right) \leq 2\left|\langle v, x_j \rangle\right|,$$

so the conclusion follows from the bounded difference inequality, Theorem 3 (i). ∎

Lemma 14. *For $v_1, \ldots, v_K \in H$ satisfying $\|v_k\| \leq 1$, $\mathbf{x} \in H^m$ we have*

$$\mathbb{E} \max_k \left| \left\langle v_k, \sum_i \sigma_i x_i \right\rangle \right| \leq \sqrt{2m \left\| \hat{\Sigma}(\mathbf{x}) \right\|_\infty} \left(2 + \sqrt{\ln K}\right).$$

Proof. Let $F_k = |\langle v_k, \sum_i \sigma_i x_i \rangle|$. Using integration by parts we have for $\delta \geq 0$

$$\begin{aligned}
\mathbb{E} \max_k F_k &\leq \sqrt{m \left\| \hat{\Sigma}(\mathbf{x}) \right\|_\infty} + \delta + \int_{\sqrt{m \left\| \hat{\Sigma}(\mathbf{x}) \right\|_\infty} + \delta}^{\infty} \max_k \Pr\{F_k \geq s\} \, ds \\
&\leq \sqrt{m \left\| \hat{\Sigma}(\mathbf{x}) \right\|_\infty} + \delta + \sum_k \int_\delta^\infty \Pr\{F_k \geq \mathbb{E} F_k + s\} \, ds \\
&\leq \sqrt{m \left\| \hat{\Sigma}(\mathbf{x}) \right\|_\infty} + \delta + \sum_k \int_\delta^\infty \exp\left(\frac{-s^2}{2m \left\| \hat{\Sigma}(\mathbf{x}) \right\|_\infty} \right) ds \\
&\leq \sqrt{m \left\| \hat{\Sigma}(\mathbf{x}) \right\|_\infty} + \delta + \frac{mK \left\| \hat{\Sigma}(\mathbf{x}) \right\|_\infty}{\delta} \exp\left(\frac{-\delta^2}{2m \left\| \hat{\Sigma}(\mathbf{x}) \right\|_\infty} \right).
\end{aligned}$$

Above the first inequality is trivial, the second follows from Lemma 13 (i) and a union bound, the third inequality follows from Lemma 13 (ii) and the last from a well known approximation. The conclusion follows from substitution of $\delta = \sqrt{2m \left\| \hat{\Sigma}(\mathbf{x}) \right\|_\infty \ln(eK)}$. ■

Proposition 15. *Let* $S_{\mathcal{E}} := \mathbb{E}_{\tau \sim \mathcal{E}} \mathbb{E}_{(\mathbf{x},\mathbf{y}) \sim \mu_\tau^m} \left\| \hat{\Sigma}(\mathbf{x}) \right\|_\infty$. *With probability at least* $1 - \delta$ *in the multisample* $\mathbf{Z} \sim \rho_{\mathcal{E}}^T$

$$\begin{aligned}
&\sup_{D \in \mathcal{D}_K} R_{\mathcal{E}}(A_D) - \frac{1}{T} \sum_{t=1}^T \hat{R}_D(\mathbf{z}_t) \\
&\leq L\alpha K \sqrt{\frac{2\pi S_1(\mathbf{X})}{T}} + 4L\alpha \sqrt{\frac{S_\infty(\mathcal{E})(2 + \ln K)}{m}} + \sqrt{\frac{\ln 1/\delta}{2T}}.
\end{aligned}$$

Proof. Following our strategy we write (abbreviating $\rho = \rho_{\mathcal{E}}$)

$$\begin{aligned}
&\sup_{D \in \mathcal{D}_K} R_{\mathcal{E}}(A_D) - \frac{1}{T} \sum_{t=1}^T \hat{R}_D(\mathbf{z}_t) \\
\leq &\sup_{D \in \mathcal{D}_K} \mathbb{E}_{\tau \sim \mathcal{E}} \mathbb{E}_{\mathbf{z} \sim \mu_\tau^m} \left[\mathbb{E}_{(x,y) \sim \mu_\tau} \left[\ell\left(\langle A_D(\mathbf{z}), x \rangle, y \right) \right] - \hat{R}_D(\mathbf{z}) \right] \\
+ &\sup_{D \in \mathcal{D}_K} \mathbb{E}_{\mathbf{z} \sim \rho} \left[\hat{R}_D(\mathbf{z}) \right] - \frac{1}{T} \sum_{t=1}^T \hat{R}_D(\mathbf{z}_t)
\end{aligned}$$

and proceed by bounding each of the two terms in turn.

For any fixed dictionary D and any measure μ on $\mathcal{Z}$ we have

$$\begin{aligned}
& \mathbb{E}_{\mathbf{z}\sim\mu^m}\left[\mathbb{E}_{(x,y)\sim\mu}\left[\ell\left(\left\langle A_D\left(\mathbf{z}\right),x\right\rangle,y\right)\right]-\hat{R}_D\left(\mathbf{z}\right)\right] \\
\leq\ & \mathbb{E}_{\mathbf{z}\sim\mu^m}\sup_{\gamma\in\mathcal{C}_\alpha}\left[\mathbb{E}_{(x,y)\sim\mu}\left[\ell\left(\left\langle D\gamma,x\right\rangle,y\right)\right]-\frac{1}{m}\sum_{i=1}^{m}\ell\left(\left\langle D\gamma,x_i\right\rangle,y_i\right)\right] \\
\leq\ & \frac{2}{m}\mathbb{E}_{\mathbf{z}\sim\mu^m}\mathbb{E}_\sigma\sup_{\gamma\in\mathcal{C}_\alpha}\sum_{i=1}^{m}\sigma_i\ell\left(\left\langle D\gamma,x_i\right\rangle,y_i\right) \text{ by Theorem 4} \\
\leq\ & \frac{2L}{m}\mathbb{E}_{\mathbf{z}\sim\mu^m}\mathbb{E}_\sigma\sup_{\gamma\in\mathcal{C}_\alpha}\sum_{k}\gamma_k\left\langle De_k,\sum_{i=1}^{m}\sigma_i x_i\right\rangle \text{ by Lemma 7} \\
\leq\ & \frac{2L\alpha}{m}\mathbb{E}_{\mathbf{z}\sim\mu^m}\mathbb{E}_\sigma\max_{k}\left|\left\langle De_k,\sum_{i=1}^{m}\sigma_i x_i\right\rangle\right| \text{ by Hölder's inequality} \\
\leq\ & \frac{2L\alpha}{m}\mathbb{E}_{\mathbf{z}\sim\mu^m}\sqrt{2m\lambda_{\max}\left(\hat{\Sigma}\left(\mathbf{x}\right)\right)}\left(2+\sqrt{\ln K}\right) \text{ by Lemma 13 (i)} \\
\leq\ & 2L\alpha\sqrt{\frac{4\mathbb{E}_{\mathbf{z}\sim\mu^m}\lambda_{\max}\left(\hat{\Sigma}\left(\mathbf{x}\right)\right)\left(2+\ln K\right)}{m}} \text{ by Jensen's inequality.}
\end{aligned}$$

This gives the bound

$$\mathbb{E}_{\mathbf{z}\sim\mu^m}\left[\mathbb{E}_{(x,y)\sim\mu}\left[\ell\left(\left\langle A_D\left(\mathbf{z}\right),x\right\rangle,y\right)\right]-\hat{R}_D\left(\mathbf{z}\right)\right]\leq 4L\alpha\sqrt{\frac{\mathbb{E}_{\mathbf{z}\sim\mu^m}\lambda_{\max}\left(\hat{\Sigma}\left(\mathbf{x}\right)\right)\left(2+\ln K\right)}{m}}$$

valid for every measure μ on $H\times\mathbb{R}$ and every $D\in\mathcal{D}_K$. Replacing μ by μ_τ, taking the expectation as $\tau\sim\mathcal{E}$ and using Jensen's inequality bounds the first term on the right hand side of (B.2) by the second term on the right hand side of (B.1).

We proceed to bound the second term. From Corollary 6 and Lemma 8 we get that with probability at least $1-\delta$ in $\mathbf{Z}\sim\left(\rho_{\mathcal{E}}\right)^T$

$$\sup_{D\in\mathcal{D}_K}\mathbb{E}_{\mathbf{z}\sim\rho}\left[\hat{R}_D\left(\mathbf{z}\right)\right]-\frac{1}{T}\sum_{t=1}^{T}\hat{R}_D\left(\mathbf{z}_t\right)\leq\frac{\sqrt{2\pi}}{T}\mathbb{E}_\zeta\sup_{D\in\mathcal{D}_K}\sum_{t=1}^{T}\zeta_t\hat{R}_D\left(\mathbf{z}_t\right)+\sqrt{\frac{\ln 1/\delta}{2T}},$$

where ζ_t is an orthogaussian sequence. Define two Gaussian processes Ω and Ξ indexed by $\mathcal{D}_K$ as

$$\Omega_D=\sum_{t=1}^{T}\zeta_t\hat{R}_D\left(\mathbf{z}_t\right)\quad\text{ and }\quad\Xi_D=\frac{L\alpha}{\sqrt{m}}\sum_{t=1}^{T}\sum_{i=1}^{m}\sum_{k=1}^{K}\zeta_{kij}\left\langle De_k,x_{ti}\right\rangle,$$

where the ζ_{ijk} are also orthogaussian. Then for $D_1, D_2 \in \mathcal{D}_K$

$$\begin{aligned}
\mathbb{E}\left(\Omega_{D_1} - \Omega_{D_2}\right)^2 &= \sum_{t=1}^{T}\left(\hat{R}_{D_1}\left(\mathbf{z}_t\right) - \hat{R}_{D_2}\left(\mathbf{z}_t\right)\right)^2 \\
&\leq \sum_{t=1}^{T}\left(\sup_{\gamma\in\mathcal{C}_\alpha}\frac{1}{m}\sum_{i=1}^{m}\ell\left(\left\langle D_1\gamma, x_{ti}\right\rangle, y_{ti}\right) - \ell\left(\left\langle D_2\gamma, x_{ti}\right\rangle, y_{ti}\right)\right)^2 \\
&\leq L^2\sum_{t=1}^{T}\sup_{\gamma\in\mathcal{C}_\alpha}\left(\frac{1}{m}\sum_{i=1}^{m}\left\langle\gamma, \left(D_1^* - D_2^*\right)x_{ti}\right\rangle\right)^2 \text{ Lipschitz} \\
&\leq \frac{L^2}{m}\sum_{t=1}^{T}\sup_{\gamma\in\mathcal{C}_\alpha}\sum_{i=1}^{m}\left\langle\gamma, \left(D_1^* - D_2^*\right)x_{ti}\right\rangle^2 \text{ Jensen} \\
&\leq \frac{L^2\alpha^2}{m}\sum_{t=1}^{T}\sum_{i=1}^{m}\sum_{k=1}^{K}\left\|\left(D_1^* - D_2^*\right)x_{ti}\right\|^2 \text{ Cauchy Schwarz} \\
&= \frac{L^2\alpha^2}{m}\sum_{t=1}^{T}\sum_{i=1}^{m}\sum_{k=1}^{K}\left(\left\langle D_1e_k, x_{ti}\right\rangle - \left\langle D_2e_k, x_{ti}\right\rangle\right)^2 = \mathbb{E}\left(\Xi_{D_1} - \Xi_{D_2}\right)^2.
\end{aligned}$$

So by Slepian's Lemma

$$\begin{aligned}
\mathbb{E}\sup_{D\in\mathcal{D}_K}\sum_{t=1}^{T}\zeta_j\hat{R}_D\left(\mathbf{z}_t\right) &= \mathbb{E}\sup_{D\in\mathcal{D}_K}\Omega_D \leq \mathbb{E}\sup_{D\in\mathcal{D}}\Xi_D \\
&= \frac{2\pi}{T}\frac{L\alpha}{\sqrt{m}}\mathbb{E}\sup_{D\in\mathcal{D}_K}\sum_{t=1}^{T}\sum_{i=1}^{m}\sum_{k=1}^{K}\zeta_{kij}\left\langle De_k, x_{ti}\right\rangle \\
&= \frac{L\alpha}{\sqrt{m}}\mathbb{E}\sup_{D\in\mathcal{D}_K}\sum_{k=1}^{K}\left\langle De_k, \sum_{t=1}^{T}\sum_{i=1}^{m}\zeta_{kij}x_{ti}\right\rangle \\
&\leq \frac{L\alpha}{\sqrt{m}}\sup_{D\in\mathcal{D}_K}\left(\sum_k\left\|De_k\right\|^2\right)^{1/2}\mathbb{E}_\zeta\left(\sum_k\left\|\sum_{t,i}\zeta_{tki}x_{ti}\right\|^2\right)^{1/2} \\
&\leq \frac{L\alpha\sqrt{K}}{\sqrt{m}}\left(\sum_k\mathbb{E}_\zeta\left\|\sum_{t,i}\zeta_{tki}x_{ti}\right\|^2\right)^{1/2} \\
&\leq \frac{L\alpha\sqrt{K}}{\sqrt{m}}\left(\sum_k\sum_{t,i}\left\|x_{ti}\right\|^2\right)^{1/2} \leq L\alpha K\sqrt{mTS_1\left(\mathbf{X}\right)}.
\end{aligned}$$

We therefore have that with probability at least $1-\delta$ in the draw of the multi sample $\mathbf{Z}\sim\rho^T$

$$\sup_{D\in\mathcal{D}_K}\mathbb{E}_{\mathbf{z}\sim\rho}\left[\hat{R}_D\left(\mathbf{z}\right)\right] - \frac{1}{T}\sum_{i=1}^{T}\hat{R}_D\left(\mathbf{Z}_{t_j}\right) \leq L\alpha K\sqrt{\frac{2\pi S_1\left(\mathbf{X}\right)}{\sqrt{T}}} + \sqrt{\frac{9\ln 2/\delta}{2T}}.$$

which in (B.2) combines with (B.2) to give the conclusion. ■

Proof of Theorem 2. Let D_{opt} and γ_τ the minimizers in the definition of R_{opt}, so that

$$R_{\mathrm{opt}} = \mathbb{E}_{\tau\sim\mathcal{E}}\mathbb{E}_{(x,y)\sim\mu_\tau}\ell\left[\left(\left\langle D_{\mathrm{opt}}\gamma_\tau, x\right\rangle, y\right)\right].$$

$R_{\mathcal{E}}\left(A_{D(\mathbf{Z})}\right) - R_{\mathrm{opt}}$ can be decomposed as the sum of four terms,

$$\begin{aligned}
&\left(R_{\mathcal{E}}\left(A_{D(\mathbf{Z})}\right) - \frac{1}{T}\sum_{t=1}^{T}\hat{R}_{D(\mathbf{Z})}\left(\mathbf{z}_t\right)\right)\\
&+\left(\frac{1}{T}\sum_{t=1}^{T}\hat{R}_{D(\mathbf{Z})}\left(\mathbf{z}_t\right) - \frac{1}{T}\sum_{t=1}^{T}\hat{R}_{D_{\mathrm{opt}}}\left(\mathbf{z}_t\right)\right)\\
&+\frac{1}{T}\sum_{t=1}^{T}\hat{R}_{D_{\mathrm{opt}}}\left(\mathbf{z}_t\right) - \mathbb{E}_{\mathbf{z}\sim\boldsymbol{\rho}}\hat{R}_{D_{\mathrm{opt}}}\left(\mathbf{z}\right)\\
&+\mathbb{E}_{\tau\sim\mathcal{E}}\left[\mathbb{E}_{\mathbf{z}\sim\mu_\tau^m}\hat{R}_{D_{\mathrm{opt}}}\left(\mathbf{z}\right) - \mathbb{E}_{(x,y)\sim\mu_\tau}\left[\ell\left(\left\langle D_{\mathrm{opt}}\gamma_\tau, x\right\rangle, y\right)\right]\right].
\end{aligned}$$

By definition of $\hat{R}$ we have for every τ that

$$\begin{aligned}
\mathbb{E}_{\mathbf{z}\sim\mu_\tau^m}\hat{R}_{D_{\mathrm{opt}}}\left(\mathbf{z}\right) &= \mathbb{E}_{\mathbf{z}\sim\mu_\tau^m}\min_{\gamma\in\mathcal{C}_\alpha}\frac{1}{m}\sum_{i=1}^{m}\ell\left[\left(\left\langle D_{\mathrm{opt}}\gamma, x_i\right\rangle, y_i\right)\right]\\
&\leq \mathbb{E}_{\mathbf{z}\sim\mu_\tau^m}\frac{1}{m}\sum_{i=1}^{m}\ell\left[\left(\left\langle D_{\mathrm{opt}}\gamma_\tau, x_i\right\rangle, y_i\right)\right] = \mathbb{E}_{(x,y)\sim\mu_\tau}\ell\left[\left(\left\langle D_{\mathrm{opt}}\gamma_\tau, x\right\rangle, y\right)\right].
\end{aligned}$$

The term (B.6) above is therefore non-positive. By Hoeffding's inequality the term (B.5) is less than $\sqrt{\ln\left(2/\delta\right)/2T}$ with probability at least $1-\delta/2$. The term B.4 is non-positive by the definition of $D\left(\mathbf{Z}\right)$. Finally we use Proposition 15 to obtain with probability at least $1-\delta/2$ that

$$\begin{aligned}
R_{\mathcal{E}}\left(A_{D(\mathbf{Z})}\right) - \frac{1}{T}\sum_{t=1}^{T}\hat{R}_{D(\mathbf{Z})}\left(\mathbf{z}_t\right) &\leq \sup_{D\in\mathcal{D}_K} R_{\mathcal{E}}\left(A_D\right) - \frac{1}{T}\sum_{t=1}^{T}\hat{R}_D\left(\mathbf{z}_t\right)\\
&\leq L\alpha K\sqrt{\frac{2\pi S_1\left(\mathbf{X}\right)}{T}} + 4L\alpha\sqrt{\frac{S_\infty\left(\mathcal{E}\right)\left(2+\ln K\right)}{m}} + \sqrt{\frac{9\ln 4/\delta}{2T}}.
\end{aligned}$$

Combining these estimates on (B.3), (B.4), (B.5) and (B.6) in a union bound gives the conclusion. ∎